

CAAS 1976 Conference
Université Laval

Clarence S. Bayne
Quantitative Methods
Concordia University
Montreal, Quebec

A Sampling Technique for the Location of a Very Rare Population

My introduction to the problem of analysis of a very rare population arises out of repeated requests to me as the Director of Research at the Black Studies Center of Montreal for social and economic indicators describing the state of existence of the Black Population in Montreal. Such an analysis as a minimum would require information about the population of the following type: the number of persons per room for Black households, the quality of housing available to Blacks, the degree of functional literacy, school retention rates, infant mortality rate, health facilities and services available to this population and income per household. A preliminary search of census statistics quickly indicated that this information could not be located in census tracks and that one was faced with a special case of a very rare population sampling problem. Immigration statistics showing country of origin of immigrants plus other statistical sources suggest that the Black population of Canada is probably about 1 percent of the total Canadian population, with major concentrations in Ontario, Quebec and Nova Scotia. The Canadian Census does not list households according to racial characteristics but rather according to ethnic and cultural characteristics and for Canadian census purposes, Black is neither considered ethnic nor a socially accepted method of classification. As a result of this definitional problem, the Black population, quite apart from the effect of its small representation in the total population at any geographic level, is further lost in the general categories of Canadian, British, American, South American, etc. There is, therefore, not only the "needle in the haystack" effect but a falsification effect operating on the population. The second effect, as shall be seen later, makes the problem of identification and classification of information very difficult.

A search of the literature on rare population sampling beginning with Hansen, Hurwitz and Madow (1), Kish (2), Deming (3) and Sudman (4), and West (5) suggests the use of a cluster sampling technique using census units as the cluster boundaries and applying a sequential sampling technique to quickly identify clusters with zero or few members of the rare population. The sequential technique makes it possible to eliminate clusters which give little or no information relative to the cost of sampling and in general require much smaller samples than for fixed sampling schemes for fixed error levels. (Wald (6) and the Columbia Statistical Research Group (7)). The sequential sampling technique by allowing the analyst to select those clusters with the greatest amount of information per dollar spent on sampling and avoid those with minimum information improves the efficiency of the cluster sampling technique used to select the general sample.

Following the Wald sequential sampling technique, one can regard screening for a very rare population in a cluster as similar to inspection for defectives in a product by sampling lots. According to Wald's technique, inspection of a census sub-division (cluster) for the rare population continues as long as:

$$\log \frac{\beta}{1-\alpha} < \log \frac{P_{1m}}{P_{0m}} < \log \frac{1-\beta}{\alpha} \quad (1)$$

where α and β are Type I and Type II and P_0 and P_1 represent different probabilities of finding a member of the rare population. Inspection is terminated at the instant that (1) does not hold. If at this stage:

$$\log \frac{P_{1m}}{P_{0m}} \geq \log \frac{1-\beta}{\alpha} \quad (2)$$

the sub-division is omitted, and if:

$$\log \frac{P_{1m}}{P_{0m}} \leq \log \frac{\beta}{1-\alpha} \quad (3)$$

the sub-division is included. The technique provides minimum strata sample sizes (h_0/s) which can be used in determining optimum strata samples when enumeration costs are known.

While sequential sampling techniques minimize the cost of selecting the sample it creates other problems, and does not solve other problems which exist in spite of it. First, there are problems of misclassification due to errors Types I and II. But these can be controlled. Another type of problem arises during the screening process that is of a more serious nature. There is the "gatekeeper" error caused by either the respondent falsifying a response to stop the interview thus causing incorrect classification, or by the interviewer omitting respondents incorrectly because of personal biases or due to weaknesses in the screening procedures. Improved field procedures for screening can eliminate many of these errors. Another type of error occurs when the respondent is identified as a member of the rare population but falsifies information pertaining to the characteristics the analyst is attempting to measure. These errors are very severe since they cannot be easily controlled by the existing sampling techniques.

In the case of this rare population the process of falsification of information was considered to be very active. Previous attempts by community groups at very limited house to house screening had resulted in a very large percentage of households stopping interviews after the stage of being identified as Black. Other households preferred to be classified under other categories although, in the opinion of the interviewer, they fell within that category. There was also reason to believe that members of this rare population exhibited a tendency to respond to questions pertaining to their life condition according to some perceived norm rather than that which actually applied to them. For these reasons and the fact that the meagre resources available for this research did not permit extensive screening of the population alternative methods had to be found that would permit a more objective

analysis, while keeping within the constraints set by limited research resources allocated to the project.

Methodology

First, an indirect screening approach was used in order to identify the location of the population within certain geographic units. The geographic unit was selected for analysis according to its importance in terms of the percentage of the very rare population concentrated in that unit. Once these regions or districts are identified and grouped to correspond to census subdivisions, a social infrastructural index for these areas as developed by Taylor (8) is applied to the particular districts. It is assumed that there is a one-to-one linear relationship between the Social Infrastructural Index (Socindex) pertaining to a particular subdivision and the set of household units of the rare population located in that geographic unit. It is further assumed that the Socindex for a geographic unit is a mirror image of the social well-being of the rare population concentrated in it. Implicit in this assumption is the further assumption that the units of rare population have equal access to or participate on an equal basis in the use of health, education, and housing facilities. This is a very feasible assumption with respect to a particular geographic unit. Any barriers to access will be reflected in terms of concentrations in regions with low as opposed to high Socindex.

It is well known that income and social stratification tends to determine where one lives and the quality of services to which one has access. In a more general sense, there is a certain functional relationship between the physical and social infrastructures of the geographic unit in which people live and their general well being. The Socindex measures this; and does not have to depend on the target population acting as the respondent to supply the information. Nor does it have to depend on an interviewer's "good" judgement in screening the rare population. Thus the use of such an index eliminates the errors outlined earlier. The major problem left therefore is to find an alternative but objective method of identifying the very rare population and locating it within particular regions.

An Indirect Screening Technique

The problem of sampling a very rare population derives largely from its lack of visibility. The members of some populations (especially organic populations) may however generate events which are easily identifiable and which are, therefore, helpful in locating the parent population through a process of indirect screening; that is, by conducting a survey of the "offspring" event and tracing it back to the parent population. For instance, since members of black household units work, then one can conduct a survey of all business enterprises to determine the number of black employees, listing them by name and address. This would be one way of locating the rare population within distinct geographic units. Similarly, household units with children of school age must according to Canadian law send these children to schools within the school district in which they live. A survey of black students in schools in Montreal, (PSBGM and MCSC) can be used as a good indicator of the location of black household units (assuming a certain ghetto effect.)

A Survey of Black Children in Montreal School System: Elementary and High Schools

As Director of Research (Black Studies Center), I conducted a survey of Black children in Montreal schools (elementary and high school) during the school sessions 1973-1974 and 1974-1975. The questionnaires were set up in such a way that the teachers or principals identify the target population according to certain characteristics. In this way, falsification on the part of the target population was avoided. At the same time the questionnaire was structured to minimize bias on the part of the enumerator. For this survey, Black was defined as one of the obvious African stock, or one of African and any other stock. The enumerator was required to record these two categories separately in order to avoid confusion or resistance to classifying mixtures as Blacks. The questionnaires also required that the members of the population be identified according to country of birth. This served as a further check. Estimates for 1974-75 are presented below according to school districts.

Table 1

Survey of Black Children in All Elementary and High Schools in Montreal, 1975

<u>District Number and Description</u>	<u>Protestant</u>	<u>French Catholic</u>	<u>English Catholic</u>	<u>Total</u>	<u>Percentage Distribution</u>
1. Dollard des Ormeaux, Pierrefonds & St. Laurent	344	85	-	429	10
2. TMR, Outremont & Eastern Mtl.	509	93	214	816	18
3. Verdun, LaSalle and Westmount	1,226	72	-	1,298	29
4. Western NDG, Cote St. Luc Lachine and Dorval	364	124	-	488	11
5. Inner-City, Eastern and N.E. fringes of NDG.	917	172	350	1,439	32
	3,360	546	564	4,470	100

The distribution of the school population is consistent with qualitative estimates of the distribution of the total Black population in Montreal as well as a priori expectations about where certain immigrant populations could afford to live.

The major concentrations of Black population are in Verdun, La Salle municipalities and the fringes of lower Westmount; NDG; Central and Eastern Montreal. If one applies a relevant Socindex to these areas then one can make objective inferences about the rare population.

Discussion of the Socindex

Theoretically the Socindex should consist of relevant social infrastructural measures (W). These should be weighted for each geographic unit (census sub-division) to derive the quality of life indicator (W_I). The weights (A_{ij}) reflect the relative importance which society as a whole places on various areas of social concern, i.e., the "revealed preferences" of society. The quality of life indicator can be written as:

$$W_I = \frac{\sum_{j=1}^n A_{ij} \cdot W_{ij}}{\sum_{j=1}^n A_{ij}}$$

Where A_{ij} is the appropriate weight accorded to the i th location and the j th infrastructural measure, W . In practice Taylor is forced to use statistical infrastructural indicators since he cannot easily determine unique measures of A_{ij} . These indicators are standardized and unit weights with appropriate signs (+) used in constructing the Socindex.

Taylor's Socindex identifies three major areas of concern, Housing, Education, and Health, all of which are highly correlated with income per capita. For each major area he selected two indicators. For houses, there are (1) persons per room, and (2) the percentage of dwellings with exclusive use of indoor flush toilet facilities; for education, there are (1) the percentage of the population aged 15 years and over with grade 9 education or more and (2) the percentage of the population 5 to 14 years old attending school full time; for health, the indicators are (1) the health personnel per capita and (2) infant mortality Z-scores.

The means and standard deviations were computed from the data on the six variables chosen. (i.e., mean and standard deviation for Canada. The Z-scores for each C.S.D. (municipality) were then computed on the basis of the following formula:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_J}{SD_J}$$

where X_{ij} is the individual value of the j th indicator in the i th census division, and $\bar{X}_J$ and SD_J are the Canada mean and standard deviation respectively. All the properties of the normal distribution are assumed to apply.

Once the Z-scores for each indicator for a particular region is calculated, the Socindex can be derived by multiplying the score by 1 or -1 depending on whether the score indicates a positive or negative contribution to wellbeing and then summed. A low valued index represents less wellbeing than a higher value index.

Application of Socindex

There was some difficulty in marrying school districts to census subdivisions so that some qualitative adjustment are necessary. Socindex obtained from Taylor's calculations are given below according to school districts and the census subdivisions that roughly fall within these districts.

TABLE 2
Socindex by School District and Census Subdivision

Districts	Index
<u>District 1</u>	<u>Socindex</u>
Dollard des Ormeaux	.86
Pierrefonds	.62
St. Laurent	.73
<u>District 2</u>	
Town of Mount Royal	1.44
Outremont	.93
Eastern Montreal	.16
<u>District 3</u>	
Verdun	.37
LaSalle	.46
Westmount	1.42
<u>District 4</u>	
Western NDG	-
Cote St. Luc	.99
Lachine	.44
Dorval	.77
<u>District 5</u>	
Montreal	.40

In applying the socindex to the black population, one must keep in mind that in District 2 the concentration of blacks in the Town of Mount Royal is limited to the Mount Royal area south of the CNR tracks close to the Outremont and Cote des Neiges area. The socindex of 1.44 is, therefore, not applicable to blacks in District 2. The index is more likely to be one varying between .16 and .40 (the Montreal index). About 40 percent of the black population in that district live in the East end of the city and is most likely French of Haitian origin. As such the index of .16 would be more applicable. On the other hand, the other 60 percent population which is largely English speaking live on the fringes of Outremont, Cote des Neige, and Mount Royal (as described above) and therefore an index between .16 and .40 is more

applicable to them. In District 3, blacks are mainly concentrated in Verdun, Lasalle and Westmount between Sherbrooke and St. Jacques (mostly around St. Antoine, St. James, Victoria, and Atwater.) The socindex of 1.42 is therefore not applicable. The Verdun Lasalle index (.37 to .46) can be generalized to the entire black population in this District. Similarly blacks living in District 4 are located largely in Lachine and to a lesser extent in Dorval therefore the relevant socindex is closer to .44 than .77. District 5 represents central Montreal and most of Eastern NDG. The concentration of blacks in this area is large as indicated by the school concentration ratios. The socindex applying to this district is .40. Last but not least District 1 represents an area to which upper income blacks migrate from other areas in the city. According to the School concentration ratios the Black population in this area is small. The Socindex applicable varies between .63 and .86.

One can now rate the Black population of Quebec according to the economic and social infrastructural class that they occupy on a scale ranging from severely disadvantaged to superior advantaged. The Taylor's index set up categories according to the following ranking procedure.

Table 3: Ranking Procedure for Socindex

Rank	Rank Description	Z-score Interval
1	Superior	greater than .750
2	Above average	0.253 to 0.750
3	Average	-0.253 to 0.253
4	Disadvantaged	-0.253 to -0.750
5	Severely Disadvantaged	greater than -0.750

According to this ranking procedure, Blacks in the five school districts were ranked as shown in Table 4.

Table 4: Ranking of Sub-division Groupings

District	Rank
1	1
2	2
3	2
4	2
5	2

Conclusion

In general, one may conclude that most Blacks in Montreal enjoy above average socio-economic conditions or at least have access to social infrastructural facilities which put them in a position above the average for the country. Some Blacks, however, especially the Haitians and Blacks living in the little Burgundy area enjoy an average or worse socio-economic life style as indicated by the Socindex for Eastern Montreal (0.16).

References

1. Hansen, Morris, Hurwitz, William, and Madow, William: "SAMPLE SURVEY METHODS AND THEORY", New York: John Wiley and Sons Inc., 1953.
2. Kish, Leslie, "SURVEY SAMPLING" New York: John Wiley and Sons, Inc. 1965
3. Deming, Edward, and Keyfitz, Nathan, "THEORY OF SURVEYS TO ESTIMATE TOTAL POPULATION", United Nations World Population Conference, August 30 - September 10, 1965.
4. Sudman, Seymour, "ON SAMPLING OF A VERY RARE HUMAN POPULATIONS", Journal of American Statistical Society, 1967, (June 1972) pp. 335-339.
5. West, Eric and Anvari, Mohsen, "AN HEURISTIC FOR SAMPLING RARE POPULATIONS", (to be published).
6. Wald, Abraham, "SEQUENTIAL ANALYSIS," New York: John Wiley and Sons, Inc. 1947.
7. Statistical Research Group, Columbia University, "SEQUENTIAL ANALYSIS OF STATISTICAL DATA: APPLICATIONS" New York: Columbia University Press, 1945.
8. Taylor, Carl A., "SOCIAL CONDITIONS REPORT (SCORE)," Division of Regional Economic Expansion, Ottawa, March 1975.